

一种基于社会化媒体和社会网络结构的 混合推荐模型

周 萍, 张子柯, 章 恬, 赵方瑜

(杭州师范大学 阿里巴巴复杂科学研究中心, 杭州 311121)

摘要: 近年来,社交网络的迅速发展为在线用户之间的沟通和交流带来极大便利,为良好的信息推荐服务提供了丰富的资源,与此同时也为个性化推荐带来了更为复杂的技术挑战.本文通过自然语言处理技术获取用户在社会化媒体(新浪微博)中的个性化兴趣标签信息,应用到自行设计开发的社会化阅读应用牛赞网中.进一步地,利用用户在牛赞网中的阅读行为和社交信息,结合用户的社会化媒体兴趣,提出了一种混合推荐模型.实验基于牛赞网中的实际数据集,并与基于用户的经典协同推荐模型进行了对比,结果表明,提出的模型在推荐性能的几个指标(AUC、准确率、召回率、多样性和新颖性)上都有很大的提高.最后,通过对牛赞网中几个典型用户进行进一步的案例分析后得出,混合推荐模型的最优参数需要根据不同社会化行为的用户进行调节.

关键词: 推荐模型; 社会网络; 社会化媒体; 标签; 混合算法

中图分类号: F 293.3

文献标志码: A

Hybrid Recommendation Model Based on Social Media and Social Networks

ZHOU Ping, ZHANG Zi-ke, ZHANG Tian, ZHAO Fang-yu

(Alibaba Research Center for Complexity Sciences, Hangzhou Normal University, Hangzhou 311121, China)

Abstract: Recently, the rapid growth of social networks has provided rich contents and yet a huge challenge for recommender systems. To better uncover its underlying role in information filtering, a hybrid algorithm was proposed based on the integrated effect of social media and social networks. The social interests(so-called tags) from social media were extracted by natural language process technology; the social influence was measured based on social network analysis; and a tunable parameter to integrate those two effects was adopted to provide recommendation results. Numerical results on a real-world dataset, *Newzan*, show that the presented model outperforms the classical user-based collaborative filtering algorithm in various metrics, including AUC (area under the curve), precision, recall, diversity and novelty. Furthermore, the case studies on three typical users from *Newzan* were performed. Statistical analyses show that the optimal recommendation parameter should vary from user to user according to the degree of their social involvement. The

收稿日期: 2013-05-21

基金项目: 国家自然科学基金资助项目(11105024, 61103109, 11205040, 11205042); 浙江省钱江人才计划资助项目(QJC1302001)

第一作者: 周 萍(1992-), 女, 本科生. 研究方向: 社会化媒体. E-mail: lotuschow23@qq.com

通讯作者: 张子柯(1982-), 男, 教授. 研究方向: 复杂网络、协同标签和推荐系统. E-mail: zhangzike@gmail.com

results can provide an in-depth understanding for unifying social media to develop social applications.

Key words: *recommendation model; social networks; social media; tags; hybrid algorithm*

随着互联网技术的不断发展与成熟,海量信息资源充斥着网络世界,用户在享受丰富信息资源的同时,也为信息超载^[1]、信息过剩^[2]等问题所困扰.面对复杂多样的网络信息资源,如何从中获取所需信息成为亟需解决的问题.个性化推荐系统也因此应运而生.个性化推荐系统以用户及相关用户群的兴趣偏好数据为基础,通过推荐算法过滤冗繁信息,为目标用户推荐喜好信息,挖掘其潜在兴趣^[3].其中,推荐算法是个性化推荐系统的核心,当前的主流推荐算法有基于内容的推荐^[4]、协同过滤推荐^[5]、基于网络结构的推荐^[6]及混合推荐^[7]等.其中,协同过滤推荐算法目前应用最为广泛,一般分为基于用户的协同过滤算法^[8]、基于内容的协同过滤算法^[9]和基于模型的协同过滤算法^[10].协同过滤推荐算法通过分析用户的行为或选择的产品的相似度来互相推荐用户可能感兴趣的信息,与其它推荐算法相比,协同过滤的优势在于:a.能够处理难以进行文本结构化表示的对象(音乐、电影等);b.具有为用户挖掘潜在兴趣的能力,能为用户推荐可能感兴趣但尚未被其发现的信息^[3].

在广泛运用协同过滤算法的同时也出现了一些具有挑战性的难题,如对于操作量较少的新用户,系统难以获得用户偏好而无法为其推荐,而新进入系统的信息,则因缺少用户评分而难以被系统推荐,即所谓的冷启动问题^[11].冷启动问题直接影响到用户使用推荐系统时的体验,这也成为了推荐系统发展面临的关键问题之一.目前已有不少学者致力于冷启动问题解决方法的研究.Massa等^[12]利用用户个人信息来预测用户对每个邻居的信任值并建立信任度量模型,以此计算得到的信任值来代替协同过滤中的相似度,该方法可以提高推荐精度并有效缓解推荐系统冷启动问题.Guy等^[13]将社会化关系分为相似关系和亲密关系,通过实验发现,相对于有着高兴趣重合度的相似关系来说,来自朋友等亲密关系用户的推荐信息更容易被用户所接受,这一研究成果为解决冷启动问题提供了重要参考.Bonhard等^[14]通过实验研究发现,当用户接触一个新的推荐系统时,为推荐项目增加解释性说明(如来自哪个朋友)能有效提高用户对推荐的接纳度.类似地,Herlocker等^[15]则发现,通过建立用户认知模型来为推荐项目提供解

释,能够有效提高用户对推荐的接纳率.

除了冷启动问题,数据稀疏性是协同过滤推荐算法中影响推荐精准性的又一重要难题.随着用户和信息数量的大幅增长,用户选择和用户评分的差异导致大量推荐项目得不到用户评价,由此出现有的用户无法获得推荐,而有的项目得不到推荐的情况^[16].为解决这一问题,张光卫等^[17]提出了基于云模型的协同过滤推荐算法,在传统的基于用户的推荐算法基础上引入了云模型,提高了相似度计算的准确性,能够有效解决数据稀疏性情况下推荐精确度的问题,从而取得了较好的推荐效果.王世云等^[18]提出了一种基于网络书签的个性化信息推荐方法,依据用户收藏的Web资源之间的关系建立Web用户关系网络,利用社团过滤算法进行社团结构划分,来实现社团内基于协作过滤的个性化信息推荐和社团间基于“信息桥”的个性化信息推荐.王丽莎等^[19]将项目评分与标签信息两者结合,提出了一种新的随机游走模型(TRWR),通过用户标签标注的信息,更加准确地捕捉到用户的兴趣特征,预测用户对于项目的评分值,从而提高个性化信息推荐的结果.丛晓琪等^[20]在传统的基于用户或资源的协同过滤推荐算法上,引入了时间权重,合理利用了用户的早期有效数据和近期数据,适应了用户动态的兴趣变化,提高了推荐结果的准确率.王晟等^[21]基于贝叶斯个性化排序的微博推荐算法,一定程度上缓解了微博用户活跃程度低、数据稀疏不对称及用户兴趣动态变化等问题.张尧等^[22]提出了基于用户分类的邻居选择方法,将邻居用户分为专家型用户、可信性用户和兴趣相似用户这3类,对应不同类型用户提供不同类别的推荐项目,提高了推荐的准确度.

除此之外,在实际应用中,随着新用户的不断进入、数据量的不断增加,大大增加了系统的计算量,延长了系统的响应时间,因此,算法的可扩展性^[11]成为制约推荐系统实际应用的重要影响因素,对此,高滢等^[23]提出了结合用户等级和似然关系模型的协同过滤推荐算法,对不同等级的用户给出不同的推荐策略,提高了推荐效率及推荐结果的准确度.赵亮等^[24]将推荐算法分为维数简化和项集相似性计算这2个过程,在提高计算精度的基础上提高了计

算效率,较为有效地解决了协同过滤推荐算法的可扩展性问题。

为得到更加高效精准的推荐方法,专家学者们从不同的切入点进行了各种尝试研究.研究表明,个体所处的集体和组织的行为能为挖掘个体的兴趣偏好提供强有力的依据,社交网络就是一个很好的例子.社交网络是指人与人之间由于某种特定目的而形成相对稳定的关系网^[25],这与同一个推荐系统中,用户间通过喜欢同一类信息所形成的关系网络相似,因此,不少学者尝试将社会网络的分析理论应用到个性化推荐中来^[26].He等^[27]通过社会网络分析方法计算朋友之间的亲密度,并以此进行推荐.Donovan等^[28]利用用户或者推荐项目预测的准确性设计信任度,增加信任度后的推荐效果相比经典的协同过滤推荐有了明显的改进.蔡浩等^[29]通过建立用户信任模型,根据综合权重进行推荐,弥补了仅基于用户相似性推荐的不足,有效地降低了用户的预测评分误差,提高了推荐质量.毛进等^[30]利用社会网络分析方法计算标签权重并将其转化为用户兴趣相似度作用于个性化推荐,取得了较好的效果。

此外,作者注意到了近年来发展迅猛的社会化媒体.社会化媒体是以互联网为介质,人们彼此用来分享意见、经验和观点的平台,它以参与性、公开性、交流性、连通性、社区化及多平台等特性赋予了每个人创作并传播内容的能力^[31],其主要的形式有社交网络(SNS网络社区)、博客、微博、视频播客、音频播客、留言板(论坛)及维基等.近年来,微博更是以其随时发布、实时沟通、综合平台、相对聚合及个性鲜明等特点成为大众网民生活中不可缺少的一部分^[32].在自我表达、求异心理及人际交往等心理需要的促使下,网民在微博上发布的信息愈发带有浓厚的个性化色彩^[33],而这无疑为推荐系统捕捉用户兴趣特征带来了极大的便利.本文通过自然语言处理方法,从用户微博中提取标签信息来代表用户的兴趣特征,计算推荐项目,并结合通过社会网络分析方法得到的推荐者给予的推荐项目,两者融合以达到更好的推荐效果。

1 模型设计

提出了一种社会化媒体和社会网络结合的推荐算法,并将其应用到系统中.牛赞网是作者自主开发的个性化阅读平台.传统的推荐平台一般是根据用户在网站中的操作行为进行推荐的,如用户对项目

的评分.同样地,在牛赞网中,用户可以通过赞、斩或者攒等(如图1所示)操作行为来表达对推荐文章的意见偏好.“赞”表示用户喜欢该推荐项目;“斩”表示用户讨厌这篇推荐文章;“攒”是指用户可以将他还未看或未看完的文章保留在“攒”中,以待日后再看.与一般推荐平台所不同的是,在牛赞网中,用户可以通过上下游的关系得到推荐.文献^[34]具体阐述了上下游的产生过程.系统通过计算为用户推荐相似用户,这些相似用户称为该用户的上游,该用户则称为上游用户的下游,上游用户“赞”的文章即会传播到下游,出现在下游用户的推荐列表中.但是,仅依靠上游的推荐,其适应性是有限的.随着社会化媒体的广泛应用,作者注意到社交媒体为用户的个性化表达提供了可能,用户的个性化兴趣偏好也由此凸显出来.可见社会化媒体为推荐系统获取用户的兴趣偏好提供了极大的便利.因此,为了更好地满足用户个性化的需求,提取了牛赞用户的微博兴趣,并结合了这两者的作用,计算出用户的微博兴趣偏好得分和牛赞上游推荐得分,其推荐系统框架结构如图2所示。

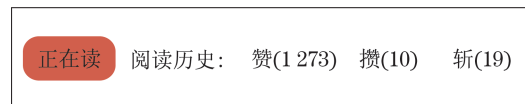


图1 牛赞用户的操作行为
Fig.1 User behaviors in Newzan

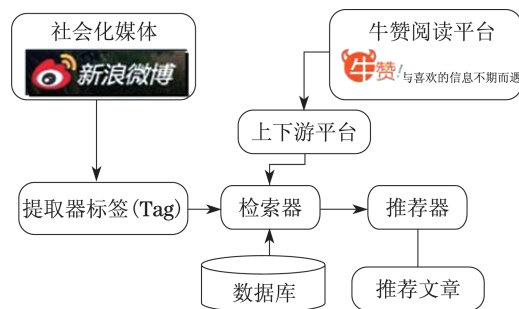


图2 推荐系统框架
Fig.2 Framework of the proposed model

如2图所示,该推荐系统结合了用户的社会化媒体兴趣标签以及其在牛赞阅读平台上的行为,形成综合权重作用于文章的推荐.具体过程如下:首先通过自然语言处理(NLP)技术^[35]与数据库中的待推荐项目进行匹配,计算出来自于微博的推荐权重;然后,对牛赞用户在网站上的行为进行分析,计算出来自用户上游的推荐得分;两者的推荐得分一起被送到推荐器中进行归一化处理,同时设置一个参数

α 值来调节微博和牛赞对推荐结果的影响权重,最终对不同用户给出符合其兴趣需求的推荐结果.

1.1 社会化媒体和牛赞上下游的推荐权重

1.1.1 用户微博偏好的推荐得分

由于微博一般由 140 字左右的汉字组成,用户的微博数量较大,如果对所有的博文进行匹配则计算工作十分繁重,所以,首先通过自然语言处理方法 TF-IDF^[36] 提取出微博标签,并计算出用户微博偏好的得分.一个项目(item)可能具备多个兴趣标签,同时,从用户微博中也可提取出多个兴趣标签,由此可见,用户通过微博标签匹配得到的推荐项目的路径有很多条,如图 3 所示.定义用户集合 $U = (u_1, u_2, u_3, \dots, u_n)$, 标签集合 $T = (t_1, t_2, t_3, \dots, t_n)$, 项目集合 $I = (i_1, i_2, i_3, \dots, i_n)$, 每个用户有其对应的标签集合 T_u . 用户的微博偏好得分

$$S_t = \sum_{t \in T_u, i \in I} \frac{1}{k_t(i)} \quad (1)$$

式中, $k_t(i)$ 为包含标签 t 的推荐项目的度.

1.1.2 用户牛赞上游的推荐得分

在牛赞网中,用户 u_i 通过其跟随的上游用户得到系统推荐,他所跟随的这些用户称为 u_i 的领导者(leader).系统通过计算用户 u_i 对推荐项目的评分,得出用户间的相似度,并以此为依据对相似用户进行排序,选取前 N 个相似用户作为用户 u_i 的上游.当用户 u_i 和用户 u_j 共同喜欢或者不喜欢的推荐项目数为 m ,不同喜好(u_i 喜欢,但 u_j 不喜欢)的项目数为 M 时,则他们的相似度可以记为 $\frac{m}{m+M}$.

考虑到这种估算方法在用户选择数目较小情况下的波动误差,为了提高相似性的精确度,采用文献[34]中的方法来计算用户(u_i, u_j)相似性,即

$$S_{ij,u} = \frac{m}{m+M} \left(1 - \frac{1}{\sqrt{m+M}} \right) \quad (2)$$

如图 4 所示,假设通过相似度计算得到用户 u_1 的下游用户为 u_2, u_3, u_4 , 用户 u_1 喜欢了文章 i , 系统自动将该篇文章推荐给他的下游 u_2, u_3, u_4 , 而用户 u_2 并不喜欢这篇文章,所以, i 并未通过 u_2 传递给他的下游,而 u_3, u_4 则将 i 传递给他们的下游,并且用户 u_7 从用户 u_3, u_4 那得到了相同的推荐项目.如果多个上游将同一个项目推荐给他们的下游,然后按照式(3)计算用户的牛赞得分

$$S_l = \sum_{l=1}^{10} \frac{1}{Nl} \quad (3)$$

式中, l 代表用户上游,牛赞中 1 个用户的上游数设

定为 10; Nl 为上游推荐给下游并被下游所喜欢的推荐项目数.

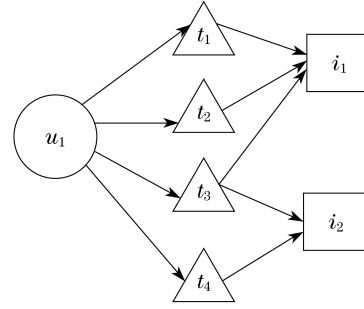


图3 用户通过标签得到推荐项目的路径

Fig.3 Paths of item recommendation getting from user's tags

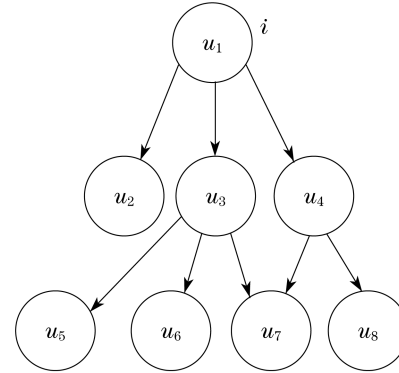


图4 用户上下游推荐机制

Fig.4 User recommendation mechanism

1.2 混合推荐模型

现引入可调参数 α 来调节用户兴趣偏好得分和用户牛赞上下游得分的权重,每个用户都有自己的最优 α 值,使得推荐结果最能体现用户的兴趣需求.由于前面的用户微博偏好权重和牛赞上下游权重并不是在统一标准下计算得出的,因此,在使用混合推荐之前,采用归一化方法将两者调整到统一标准下,如式(4)所示.

$$\left. \begin{aligned} S'_t &= \frac{S_t}{\sum S_t} \\ S'_l &= \frac{S_l}{\sum S_l} \\ S_i &= \alpha S'_t + (1 - \alpha) S'_l \end{aligned} \right\} \quad (4)$$

式中, S'_t 为归一后的用户微博偏好得分; S'_l 为归一后的用户牛赞得分; S_i 为项目 i 的最终推荐得分.

根据 S_i 值倒序排列并取得分靠前的 L 个项目推荐给用户.

2 实验及结果分析

2.1 实验数据处理

实验采用牛赞网中的用户操作数据,通过筛选得到至少赞过2条推荐项目的用户1381个,通过分词得到的标签48543条,系统推荐项目37968条.实验中随机选取数据集中90%的数据作为训练集,其余10%的数据作为测试集.

2.2 评价指标

在推荐算法的评价指标中,应用较为广泛的有准确率、召回率、AUC(area under the curve)、多样性和新颖性^[37].

准确率:指的是推荐列表中用户喜欢的推荐项目数占所有推荐项目总数的比率^[38],如式(5)所示.

$$P_u(L) = \frac{N_{tp}^L}{N_{tp}^L + N_{fp}^L}$$

$$P(L) = \frac{1}{|U|} \sum_{u \in U} P_u(L) \quad (5)$$

式中, $P_u(L)$ 为用户 u 的准确率; N_{tp}^L 表示存在于用户 u 的推荐列表中并被用户喜欢的推荐项目; N_{fp}^L 表示存在于用户 u 的推荐列表中但不被用户喜欢的推荐项目; L 代表推荐列表长度; $|U|$ 表示实验中用户的总数; $P(L)$ 为系统整体推荐准确率.

最终求均值得到系统整体推荐准确率 $P(L)$.准确率用来描述用户对于一个推荐项目感兴趣的可能性.

召回率:指推荐列表中用户喜欢的推荐项目的数目占推荐系统中用户喜欢的所有推荐项目数的比率,如式(6)所示.

$$R_u(L) = \frac{N_{tp}^L}{N_{tp}^L + N_{fn}^L}$$

$$R(L) = \frac{1}{|U|} \sum_{u \in U} R_u(L) \quad (6)$$

式中, $R_u(L)$ 为用户 u 的召回率; N_{fn}^L 代表系统并未推荐但用户喜欢的推荐项目; $R(L)$ 为系统整体召回率.

同理得系统整体召回率 $R(L)$.召回率用来描述用户喜欢的推荐项目可以被推荐系统所发现并推荐给用户的可能性.

AUC(area under the curve):对于没有明确推荐列表长度的推荐系统来说,AUC指标更能衡量推荐效果的准确性^[38].AUC指标对应ROC(receiver operator curve)曲线^[39]下的面积,AUC值越接近于1,表示推荐效果越好.

多样性:为了衡量推荐系统推荐内容的差异性,

周涛等^[40]提出了用海明距离(Hamming distance)衡量推荐列表多样性的评价指标.多样性分为用户间多样性(inter-user diversity)和用户内多样性(intra-user diversity),本文采用的是用户间多样性,如式(7)所示.

$$H_{ij}(L) = 1 - \frac{Q_{ij}}{L}$$

$$H(L) = \frac{2}{|U|} \sum H_{ij} \quad (7)$$

式中, $H_{ij}(L)$ 为用户 u_i, u_j 的海明距离; Q_{ij} 代表用户 u_i 和用户 u_j 的推荐列表中相同推荐项目的个数; $H(L)$ 为系统的海明距离.

对 $H_{ij}(L)$ 取平均值则得到了系统的海明距离 $H(L)$.用户间多样性用来描述系统对不同用户推荐不同项目的能力.

新颖性:推荐系统的新颖性是指系统为用户推荐其从未见过但是感兴趣的项目的能力,本文用推荐项目的平均流行度来衡量,流行度越低,代表推荐项目越新颖,如式(8)所示.

$$X(L) = \frac{\sum_u \sum_{i=1}^L k_u(i)}{|U|L} \quad (8)$$

式中, $k_u(i)$ 表示系统为用户 u 推荐的项目 i 的流行度,即在推荐系统中喜欢推荐项目 i 的用户数; $X(L)$ 为推荐项目的平均流行度.

2.3 实验结果与分析

实验考察了推荐列表长度 $L=10, 20, 30$ 的情况下本文所提出模型的推荐效果,同时将实验结果与经典的基于用户的协同过滤推荐算法进行了比较,如图5和图6所示(见下页). P_{AUC} 为AUC的值.

图5(a)显示,基于社会化媒体的推荐模型(即本文所提方法)的准确率总体高于经典协同过滤推荐算法的.本文所提方法,无论推荐列表长度是10,20,还是30,随着 α 值的上升,准确率都呈持续下降趋势,且 α 值越接近于1,准确率越低,这表明,牛赞上游占推荐比重越大,用户对推荐项目感兴趣的可能性越高.从图5(b)中可以看到,本文所提方法与经典协同过滤推荐算法在多样性上相差不大,但本文所提方法的推荐多样性总体仍高于经典协同过滤推荐算法的.而在本文所提方法的推荐中,无论推荐列表长度为10,20,还是30,推荐列表多样性都随 α 值平缓上升到一定高度后急速下降;且 α 值在0.5左右时,推荐的多样性最大; α 值为1时,多样性最小,推荐列表长度越小,系统为每个用户推荐的差异性越大,这表明,由微博而来的推荐差异性较小.在

图 5(c)中,本文所提方法的召回率都远高于经典协同过滤推荐算法的,最高值为经典协同过滤方法的 3 倍.而在基于社会化媒体的推荐中,无论推荐列表长度为 10,20,还是 30,随着 α 值的上升,召回率都

呈下降趋势,且 α 值越接近于 1,召回率越低,这表明,牛赞上游占推荐比重越大,用户兴趣偏好被系统发现并推荐给用户的可能性越高.图 5(d)显示,经典协同过滤中推荐项目的平均流行度均高于基于社

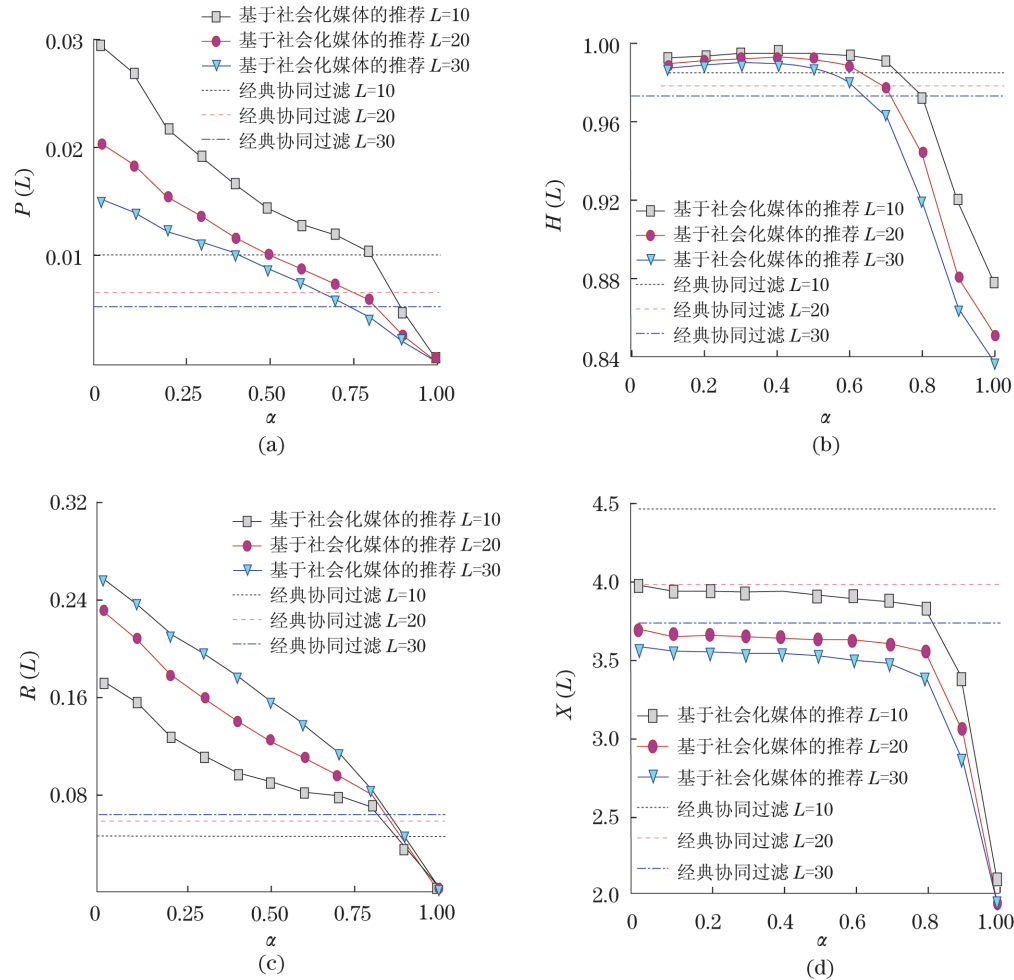


图 5 推荐列表长度为 $L=10, 20, 30$ 的情况下,基于社会化媒体的推荐与经典协同过滤推荐评价指标对比

Fig. 5 Comparison between the evaluation indices based on the proposed model (solid line) and on the classical collaboration filtering (dash line) under the list length $L=10, 20, 30$

会化媒体的推荐.而在基于社会化媒体的推荐中,无论推荐列表长度为 10,20,还是 30,随着 α 值的增大,新颖性都呈现总体下降的趋势, $\alpha=0$ 时,新颖性最高,这表明,由牛赞上游产生的推荐更能符合用户的个性化兴趣偏好,为用户提供相对冷门但符合用户兴趣的推荐项目.此外,如图 6 所示,经典协同推荐算法的 AUC 指标远低于本文所提方法的.在基于社会化媒体的推荐算法中,AUC 指标随 α 值增大而上升至接近 0.7 后开始下降. α 值在 0.1 左右,AUC 指标达到最高值; α 值为 1 时,AUC 指标值最低,由此表明,牛赞上游占推荐比重较高并结合微博提供辅助作用时,推荐系统的推荐效果最好.

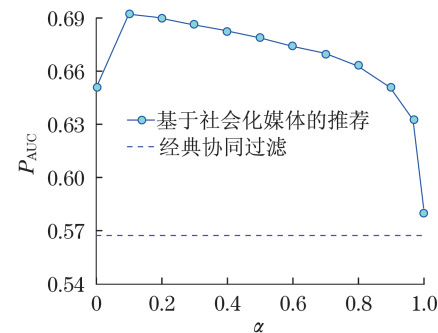


图 6 基于社会化媒体的推荐与经典协同过滤推荐 AUC 指标对比

Fig. 6 AUC vs. α . circle line: the proposed model; dash line: the classical collaboration filtering

3 案例分析

3.1 案例用户选取

以上实验计算表明,总体而言,用户在社会化媒体(微博)和推荐网站上不同的操作行为对个性化推荐结果的影响程度是不同的.而从个体而言,参数 α 的最优值越趋向于 1 的用户,社会化媒体对其推荐结果的影响越大; α 值越趋向于 0,用户在推荐平台上的操作对其推荐结果的影响更显著;而 α 的最优值介于 0~1 之间的用户,其推荐结果是在两者共同作用下更优.

为了从微观层面进一步分析用户行为之间的差异性,选取了 2011 年 9 月至 2012 年 9 月牛赞推荐平台上的用户作为案例研究的对象,将牛赞中的用户分为 3 类: α 趋向 1 的用户; α 趋向 0 的用户和 α 介于 0~1 之间的用户.通过对这些数据的分析来进一步验证实验结果.

3.2 案例数据选取和衡量标准

本文的案例分析分为以下几个步骤进行:

a. 标签选取.

首先人工选取了兴趣标签,并对数据进行分类.定义的兴趣标签包括新闻、科技、财经、人文、体育、娱乐、时尚、生活、图片、情感、音乐和其它这 12 类.其中,新闻包括时政评论、国际时事、社会动态等通讯消息;科技包含了 IT 业界、手机数码、环球地理、科学探索、生物技术等相关信息;人文涵盖了人物历史、图书文学、文化艺术等内容;娱乐包括了流行音乐、影视资讯、趣味笑话、搞笑乐图等;生活包含了医药、养生、旅游、美食等内容;情感主要是指个人感悟、情感表达、人际交往等方面;对于一些生活琐事等较为零散的内容则归类到其它标签下.

b. 数据获取.

首先通过网络爬虫技术抓取了 200 个用户的微博信息,并从数据库中提取了牛赞用户的操作数据,利用自然语言处理技术提取出用户在微博和牛赞上的操作兴趣,并将这些统计数据归类到人工标注的兴趣标签中.根据牛赞用户的分类标准,各选取了 3 类用户进行实例分析,每个标签对应的数值为标签数所占兴趣总数的百分率(括号里的数值为标签数目),如表 1 所示.其中,a 类用户的最优 α 值为 1;b 类用户的最优 α 值为 0.21;c 类用户的最优 α 值为 0.

c. 衡量标准.

表 1 a,b,c 类用户的微博和牛赞兴趣比例(数目)

Tab.1 Interests of user type a,b,c
(Microblog vs. Newzan)

兴趣 标签	$\alpha=1$,用户 a		$0<\alpha<1$,用户 b		$\alpha=0$,用户 c	
	微博	牛赞	微博	牛赞	微博	牛赞
新闻	11%(61)	14%(68)	7%(29)	7%(5)	5%(65)	7%(1)
科技	39%(216)	32%(160)	14%(60)	26%(18)	13%(154)	79%(11)
财经	10%(56)	8%(42)	4%(15)	1%(1)	2%(20)	7%(1)
人文	11%(58)	7%(36)	3%(11)	0%(0)	12%(147)	7%(1)
体育	11%(60)	9%(44)	2%(7)	6%(4)	2%(28)	0%(0)
娱乐	6%(34)	8%(40)	18%(75)	28%(19)	11%(138)	0%(0)
时尚	0%(0)	2%(10)	0%(0)	6%(4)	1%(14)	0%(0)
生活	2%(12)	3%(15)	23%(95)	3%(2)	11%(130)	0%(0)
图片	1%(6)	3%(16)	6%(24)	0%(0)	14%(169)	0%(0)
情感	4%(20)	7%(34)	15%(61)	9%(6)	12%(148)	0%(0)
音乐	4%(20)	6%(32)	3%(12)	10%(7)	8%(94)	0%(0)
其它	1%(6)	1%(4)	5%(27)	4%(2)	9%(113)	0%(0)

采用 Pearson 相关系数^[41] 指标来计算用户的微博操作和牛赞操作的相似性.定义用户微博操作兴趣标签的百分率为 W ,在牛赞中的操作兴趣标签百分率为 N ,则用户微博(W)和牛赞(N)上的相似性由 Pearson 相关系数 r 表示.

$$r = \frac{n \sum WN - \sum W \sum N}{\sqrt{n \sum W^2 - (\sum W)^2} \sqrt{n \sum N^2 - (\sum N)^2}} \quad (9)$$

式中, n 为样本个数.

根据相关系数的性质^[42]可知, $r>0$,表明存在正相关关系; $r<0$,则表明存在负相关关系.并且, $|r|\rightarrow 1$,说明 2 个变量之间的线性关系越强;反之, $|r|\rightarrow 0$,说明 2 个变量之间的线性关系越弱.一般地, $|r|\geq 0.8$,可视为高度相关; $0.5\leq |r|<0.8$,为中度相关; $0.3\leq |r|<0.5$,为低度相关; $|r|<0.3$,关系极弱,可以视为不相关.

3.3 假设检验

绘制了 3 类典型用户微博偏好兴趣和牛赞操作兴趣的散点图,并具体计算出微博和牛赞操作兴趣的 Pearson 相关系数.由于样本易受抽样波动的影响,所以,为了体现出总体的相关程度,还对样本相关系数的显著性进行了统计分析,利用 P 值进行假设检验^[42]. P 值(P value)是拒绝原假设会犯错的概率.根据小概率原理, P 值越小,拒绝原假设的理由就越充分,并且 P 值小于显著性水平,就可以拒绝原假设.假设:

H_0 :总体用户微博和牛赞兴趣的相关系数不

显著;

H1:总体用户微博和牛赞兴趣的相关系数显著.

通过 P 值检验可以确定总体数据的相关性与研究假设是否相符.

图 7(见下页)为 a,b,c 类用户微博和牛赞兴趣偏好的散点图.图 7 中包含了各用户的回归拟合方程,如实线.其中,纵轴为用户微博兴趣标签的百分率,横轴为用户牛赞兴趣标签的百分率.从图中可以看出,a 类用户($\alpha=1$)在微博和牛赞上的操作兴趣的离散程度较小,两者兴趣比率高度相似.如图 7 所

示,a 类用户在两种平台上都对科技类的信息有着较高的关注度,分别为 39% 和 32%,而对时尚、图片、生活及其它类的信息则都表现出较低的兴趣;b 类用户($0<\alpha<1$)在微博上最为关注生活类的信息,占 23% 的兴趣比率,其次是娱乐、情感类,分别为 18% 和 15%,而在牛赞平台上,其对生活类信息的关注度远不及娱乐和科技类;对比而言,c 类用户($\alpha=0$)在微博和牛赞上的操作兴趣几乎完全不同,该用户在微博中对图片、科技、情感、生活、娱乐等都有较高的关注度,而在牛赞平台上只对“科技”类信息表现出了极大的兴趣.

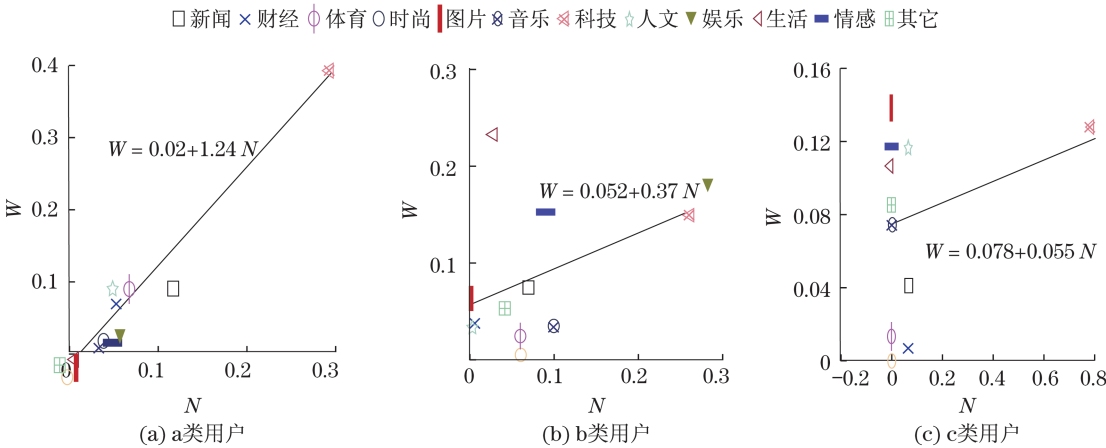


图 7 a,b,c 类用户在两种平台操作的相关性示意图(实线为线性拟合曲线)
Fig.7 Correlation index of user type a,b,c (Microblog vs. Newzan) (solid line: linear fitting)

为了进一步检验结果,由表 2 的 r 值,根据 Pearson 相关相似性的性质可以得到以下结论(其中规定显著性水平为 0.01):

表 2 a,b,c 类用户微博和牛赞兴趣偏好相似度对比
Tab.2 Interests similarity of user type a,b,c (Microblog vs. Newzan)

用户	α 值	r 值	P 值
$\alpha=1$	a	0.974	0.000
$0<\alpha<1$	b	0.472	0.122
$\alpha=0$	c	0.283	0.374

a. $\alpha=1$ 的用户,即 a 类用户,其相关系数大于 0.8,说明用户 a 的微博和牛赞的操作兴趣高度相关.由于牛赞的推荐是结合用户的微博动态以及分析用户在牛赞中的操作行为产生推荐结果的,如果用户在牛赞中的阅读兴趣和微博中的兴趣高度吻合,则可以证明微博的操作兴趣对这个用户的推荐结果有更大的影响.另外,为了检验样本相关系数的

可靠性,将该用户的 P 值与显著性水平进行比较,结果 P 值小于 0.01,通过拒绝原假设,可以相信,对于 $\alpha=1$ 的这类用户而言,微博的信息对其推荐结果有更大的作用.

b. α 介于 0~1 之间的用户,即 b 类用户,其相关系数介于 0.3~0.5 之间,可以视为低度相关,并且其 P 值大于 0.01,所以,接受原假设,可以认为 α 介于 0~1 之间的这类用户,微博的兴趣对其推荐结果并没有太大作用.

c. $\alpha=0$ 的用户,即 c 类用户的 r 值小于 0.3,其相关程度极弱,可以视为不相关.并且根据 P 值与显著性水平的比较可以看出, P 值明显大于显著性水平,所以,微博的兴趣对该类用户的推荐结果几乎没有影响,其推荐依据主要是来自用户在牛赞平台中的推荐情况.

根据以上实例的分析结果,可以验证本文的假设是合理的,微博和用户行为对不同用户的推荐结果的影响是不同的.

4 结论与展望

分析了在线用户在社会化媒体新浪微博中的个性化兴趣和其在个性化阅读平台牛赞网中的阅读偏好,同时考虑了用户在牛赞网中的阅读兴趣和社交网络结构特征,在此基础上提出了一种融合微博兴趣标签信息和牛赞网阅读上游意见的推荐模型.实验结果表明,本文所提方法无论在 AUC、准确率、召回率、多样性和新颖性这些推荐性能指标上都有一定的改进.进一步的案例分析表明,对于不同阅读习惯的用户,其最佳推荐效果对应的推荐参数会有所不同.统计分析结果显示,在设计推荐模型的时候要根据不同用户在特定的个性化环境中的融入程度来调节适当的推荐参数,以获取最佳的用户体验.

本文的主要成果体现在 4 个方面:a. 利用自然语言处理基础得到用户在社会化媒体中的兴趣标签;b. 利用社会网络分析方法获得阅读上游对用户的影响程度;c. 通过可调参数将用户在外界的兴趣喜好和阅读平台内的阅读偏好融合在一起,预测出用户平台内信息的潜在喜好程度;d. 分析得出阅读平台内不同阅读习惯和融入程度的用户有着不同的阅读偏好,需要根据个体用户的不同情况设定不同的推荐参数值.

除此之外,本文所提方法也有不足之处和需要改进的地方:a. 过滤掉了数据集中的时间戳因子,没有考虑用户的阅读兴趣特征和新闻热度会随着时间的变化而发生改变;b. 在提取微博兴趣标签时,受到分词技术的制约,没有考虑到标签之间的语义相似性及动力学机制.这些不足也成为以后改进的工作方向,作者会在本文的基础上继续探索,希望进一步利用社会化媒体丰富的资源提高个性化信息推荐的效果.

参考文献:

- [1] 李书宁. 互联网信息环境中信息超载问题研究[J]. 情报科学, 2005, 23(10): 1587 - 1590.
- [2] 杨彦辉. 信息过剩现象及其负面影响[J]. 社会科学论坛, 2001(3): 33 - 36.
- [3] 刘建国, 周涛, 汪秉宏. 个性化推荐系统的研究进展[J]. 自然科学进展, 2009, 19(1): 1 - 15.
- [4] Pazzani M J, Billsus D. Content-based recommendation systems[J]. The Adaptive Web, 2007, 4321: 325 - 341.
- [5] Herlocker J L, Konstan J A, Terveen L G, et al. Evaluating collaborative filtering recommender systems[J]. ACM Transactions on Information Systems, 2004, 22(1): 5 - 53.
- [6] Zhou T, Jiang L L, Su R Q, et al. Effect of initial configuration on network-based recommendation[J]. EPL, 2008, 81(5): 58004.
- [7] Basu C, Hirsh H, Cohen W. Recommendation as classification: using social and content-based information in recommendation[C] // Papers from 1998 Workshop, Technical Report WS - 98 - 08, Menlo Park: AAAI Press, 1998: 714 - 720.
- [8] Wang J, De Vries A P, Reinders M J T. Unifying user-based and item-based collaborative filtering approaches by similarity fusion [C] // SIGIR '06 Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: Association for Computing Machinery, 2006: 501 - 508.
- [9] Balabanovic M, Shoham Y. Fab: content-based collaborative recommendation[J]. Communications of the ACM, 1997, 40(3): 66 - 72.
- [10] Billsus D, Pazzani M J. User modeling for adaptive news access [J]. User Modeling and User-Adapted Interaction, 2000, 10(2/3): 147 - 180.
- [11] 孙冬婷, 何涛, 张福海. 推荐系统中的冷启动问题研究综述[J]. 计算机与现代化, 2012(5): 59 - 63.
- [12] Massa P, Avesani P. Trust-aware recommender systems[C] // Proceedings of the 2007 ACM Conference on Recommender Systems. New York: ACM Press, 2007: 17 - 24.
- [13] Guy I, Zwerdling N, Carmel D, et al. Personalized recommendation of social software items based on social relations [C] // Proceedings of the 3rd ACM Conference on Recommender Systems. New York: ACM Press, 2009: 53 - 60.
- [14] Bonhard P, Sasse M A. 'Knowing me, knowing you' - using profiles and social networking to improve recommender systems[J]. BT Technology Journal, 2006, 24(3): 84 - 98.
- [15] Herlocker J L, Konstan J A, Riedl J. Explaining collaborative filtering recommendations[C] // CSCW'00 Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work. New York: ACM Press, 2000: 241 - 250.
- [16] Zeng A, Yeung C H, Shang M S, et al. The reinforcing influence of recommendations on global diversification[J]. EPL, 2012, 97(1): 18005.
- [17] 张光卫, 李德毅, 李鹏, 等. 基于云模型的协同过滤推

- 荐算法[J]. 软件学报, 2007, 18(10): 2403 - 2411.
- [18] 王世云, 王志军, 李运杰. 一种基于网络书签的个性化信息推荐方法[J]. 计算机系统应用, 2011, 20(5): 180 - 183.
- [19] 王丽莎, 张绍武, 林鸿飞. 基于项目和标签的随机游走个性化信息推荐模型[J]. 情报学报, 2012, 31(3): 289 - 296.
- [20] 丛晓琪, 杨怀珍, 刘枚莲. 基于时间加权的协同过滤算法研究[J]. 计算机应用与软件, 2009, 26(8): 120 - 121.
- [21] 王晟, 王子琪, 张铭. 个性化微博推荐算法[J]. 计算机科学与探索, 2012, 6(10): 895 - 902.
- [22] 张尧, 冯玉强. 协同过滤推荐中基于用户分类的邻居选择方法[J]. 计算机应用研究, 2012, 29(11): 4216 - 4219.
- [23] 高滢, 齐红, 刘杰, 等. 结合似然关系模型和用户等级的协同过滤推荐算法[J]. 计算机研究与发展, 2008, 45(9): 1463 - 1469.
- [24] 赵亮, 胡乃静, 张守志. 个性化推荐算法设计[J]. 计算机研究与发展, 2002, 39(8): 986 - 991.
- [25] Wasserman S, Faust K. Social network analysis: methods and applications[M]. New York: Cambridge University Press, 1994.
- [26] 王国霞, 刘贺平. 个性化推荐系统综述[J]. 计算机工程与应用, 2012, 48(7): 66 - 76.
- [27] He J M, Chu W W. A social network-based recommender system (SNRS) [J]. Data Mining for Social Network Data, 2010(12): 47 - 74.
- [28] Donovan J O, Smyth B. Trust in recommender systems [C]//Proceedings of the 10th International Conference on Intelligent User Interfaces. New York: ACM Press, 2005: 167 - 174.
- [29] 蔡浩, 贾宇波, 黄成伟. 结合用户信任模型的协同过滤推荐方法研究[J]. 计算机工程与应用, 2010, 46(35): 148 - 151.
- [30] 毛进, 易明, 操玉杰, 等. 一种基于用户标签网络的个性化推荐方法[J]. 情报学报, 2012, 31(1): 23 - 30.
- [31] 王明会, 丁焰, 白良. 社会化媒体发展现状及其趋势分析[J]. 信息通信技术, 2011(5): 5 - 10.
- [32] 殷俊, 何芳. 微博在我国的传播现状及传播特征分析[J]. 河南大学学报, 2011, 51(3): 124 - 129.
- [33] 杨圣娇. 自媒体传播心理动因及个体角色重塑[J]. 现代试听, 2012(1): 32 - 36.
- [34] Medo M, Zhang Y C, Zhou T. Adaptive model for recommendation of news[J]. EPL, 2009, 88(3): 38005.
- [35] 熊回香, 夏立新. 自然语言处理技术在中文全文检索中的应用[J]. 情报理论与实践, 2008, 31(3): 432 - 435.
- [36] 张玉芳, 彭时名, 吕佳. 基于文本分类 TFIDF 方法的改进与应用[J]. 计算机工程, 2006, 32(19): 76 - 78.
- [37] 刘建国, 周涛, 郭强, 等. 个性化推荐系统评价方法综述[J]. 复杂系统与复杂性科学, 2009, 6(3): 1 - 10.
- [38] 朱郁筱, 吕琳媛. 推荐系统评价指标综述[J]. 电子科技大学学报, 2012, 41(2): 163 - 175.
- [39] Hannley J A, Mcneil B J. The meaning and use of the area under a receiver operating characteristic (ROC) curve[J]. Radiology, 1982, 143(1): 29 - 36.
- [40] Zhou T, Kuscsik Z, Liu J G, et al. Solving the apparent diversity accuracy dilemma of recommender systems [J]. Proceedings of the National Academy of Sciences of the USA, 2009, 107(10): 4511 - 4515.
- [41] Sarwar B, Karypis G, Konstan J, et al. Item-based collaborative filtering recommendation algorithms[C]//Proceedings of the 10th International Conference on World Wide Web. New York: ACM Press, 2001: 285 - 295.
- [42] 贾俊平, 何晓群, 金勇进. 统计学[M]. 4版. 北京: 中国人民大学出版社, 2009.

(编辑: 石 瑛)